From Text Summarisation to Style-Specific Summarisation for Broadcast News

From text summarisation to style-specific summarisation for broadcast news

Heidi Christensen[1], BalaKrishna Kolluru[1], Yoshihiko Gotoh[1], and Steve Renals[2]

[1] Department of Computer Science, University of Sheffield
Sheffield S1 4DP, UK
{h.christensen, b.kolluru, y.gotoh}@dcs.shef.ac.uk
[2] The Centre for Speech Technology Research, University of Edinburgh
Edinburgh EH8 9LW, UK
s.renals@ed.ac.uk

Abstract. In this paper we report on a series of experiments investigating the path from text summarisation to style-specific summarisation of spoken news stories. We show that the portability of traditional text summarisation features to broadcast news is dependent on the diffusiveness of the information in the broadcast news story. An analysis of two categories of news stories (containing only read speech or including some spontaneous speech) demonstrates the importance of the style and the quality of the transcript, when extracting the summary-worthy information content. Further experiments indicate the advantages of doing style-specific summarisation of broadcast news.

1 Introduction

A television or radio news broadcast consists of a set of stories, containing a wide variety of content, and presented in a number of styles. A broadcast news story is often a complex composition of several elements, including both planned speech (usually read) and spontaneous speech, such as a reaction or an answer.

Printed news stories typically present the most important facts in the opening line, with subsequently related facts presented in the order of decreasing importance (the "inverted information pyramid"): indeed the opening line is often referred to as the "summary lead". Broadcast news tends to be rather different: it is written to be heard, and the lead sentence(s) often aim to capture the interest of the viewer or listener, without summarising the main facts. Furthermore, the information density within the story depends on the style: for example, the news anchor may speak information-rich sentences, compared with an interviewee. This implies that the most important information, from a summarisation viewpoint, is not distributed similarly throughout all news stories. The location of regions with a high information density may depend on the style, calling for summarisation techniques that are less rigid than those typically used for the summarisation of printed news.

In this paper we present some recent work in the area of broadcast news summarisation using sentence extraction techniques. The work has been aimed at investigating the path from text-summarisation to *style-specific* summarisation of spoken news stories. We have addressed three key questions:

Q1: How well do the extractive summarisation techniques developed for text documents fare on speech? If possible, we would like to reuse textual summarisation techniques when summarising spoken language. However, speech transcripts differ from text documents in both structure and language, warranting an investigation of several issues concerning such a transfer to the speech domain. We report on a series of experiments that address the performance of individual features when applied in both text and speech summarisation, as well the effect of applying a text inspired summariser to erroneous speech recogniser transcripts (section 2). This is done by using a text corpora as well as a speech corpora, with both human ("closed-caption") and automatic speech recognition (ASR) transcripts for the broadcast TV news programmes.

Q2: To what degree is the performance of summarisers employing these text-based features dependent on the style of the broadcast news stories? There are a number of subtle differences between spontaneous and read speech [1]. Stylistically news stories with spontaneous speech tend to have the summary-worthy information distributed across the document, whereas read news stories tend to start off with a "summary lead", getting into more detail as the story progresses; adhering much more to the style of printed news stories. We have investigated the effect of the text-based features when applying a classification of news stories into two categories : stories with spontaneous elements (SPONTANEOUS) and purely read stories (READ). The analysis was carried out using the same database of spoken news stories as the first series of experiments, and effects are quantified on both closed-caption transcripts, low word error rate (WER) and high WER automatic transcripts (section 3).

*Q3: Using the established categories (*SPONTANEOUS *and* READ*), what is the observed interaction between the summarisation technique employed and the style of news story?* The automatic summarisation techniques that we have investigated are based on sentence extraction, using novelty factor, content, and context as their respective criterion for summarisation (section 4). These automatic summarisers are compared against human generated summaries. Since we are primarily concerned with the interaction between summarisation techniques and broadcast style, in these experiments, we have used hand transcribed news broadcasts, that have been manually classified to appropriate categories, so that speech recognition errors are excluded.

2 Investigating the portability of text features to the speech domain

For text it has been found that good extractive summarisers depend heavily on features relating to the content of the text [2] and on the structure and style of the text [3–5].

Content-based features are clearly vulnerable to errors introduced by a speech recognisers, and in this section we present experiments that quantify the effect of recognition errors on summarisation.

2.1 Data

Broadcast news data. For the one-sentence summarisation work we used a set of 114 ABC news broadcasts (ABC_SUM) from the TDT–2 broadcast news corpus[3], totalling 43 hours of speech. Each programme spanned 30 minutes as broadcast, reduced to around 22 minutes once advert breaks were removed, and contained on average 7–8 news stories, giving 855 stories in total. In addition to the acoustic data, both manually-generated closed-caption transcriptions and transcriptions from two different ASR systems (with high and low WERs respectively), are available [7].

All ABC_SUM transcripts have been segmented at three levels: 1) sentence boundaries (hand-segmented), 2) speaker turns (produced by LIMSI [8] for TREC/SDR) and 3) story boundaries (the individual news stories were hand-segmented as part of the TREC/SDR evaluations).

For each segmented story in the ABC_SUM data, a human summariser selected a single sentence as a "gold-standard", one-sentence extractive summary. These one-sentence summaries were all produced by the same human summariser, and validated in an evaluation experiment for their consistency and quality (see [9] for further details).

Two subsets of the data were used for training and developmental tests, containing 33.8 and 3.9 hours of speech respectively.

Newspaper data. We have used text data obtained from the DUC-2001[4] text summarisation evaluation. This data consists of newspaper stories originally used in the TREC–9 question answering track, totalling 144 files (132 for training, 12 for testing) from the Wall Street Journal, AP newswire, San Jose Mercury News, Financial Times, and LA Times, together with associated multi-line summaries[5]. Each document comprises a single news story topic, and the data is from the period 1987-1994. Although the speech data is from February to June 1998, the broad topics covered in the two data sets are very similar.

[3] The TDT–2 [6] corpus was used in the NIST Topic Detection and Tracking (TDT) evaluations and in the TREC–8 and TREC–9 spoken document retrieval (SDR) evaluations. The one-sentence summaries for the ABC_SUM data were generated at University of Sheffield

[4] url = http://www-nlpir.nist.gov/projects/duc/index.html

[5] Extractive summaries for this data were contributed by John Conroy (IDA) as an addition to the non-extractive summaries distributed with the original DUC-2001 data, and were derived to cover the same content as the non-extractive summaries.

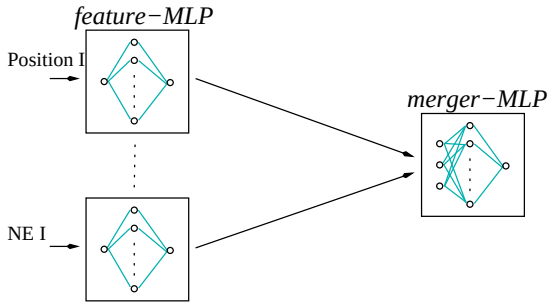

Fig. 1. Summariser architecture. All MLPs used in this work had 20 hidden units in a single hidden layer.

2.2 Summarisation approach

The summarisation task is to automatically generate an extractive summary for a spoken or printed news story. Our approach uses a trainable, feature-based model which assigns a score to each sentence that indicates how suitable that sentence is for inclusion in a summary. When generating an N-line summary, the summary is comprised of the N highest-scoring sentences.

A set of features are extracted for each sentence. The summariser is based around a set of multi-layer perceptron (MLP) classifiers [10]:, one for each feature (*feature-MLPs*) and a second level MLP (*merger-MLP*) which combines the outputs of the *feature-MLPs* (figure 1). This feature-based approach is somewhat similar to that employed by [11]; that approach discretised the features and was based on a Naive Bayes classifier. The training set for each *feature-MLP* consists of a set of single feature inputs, together with the summarisation label from the gold-standard (1 or 0), for each sentence. Thus each *feature-MLP* is trained to optimise summarisation for that feature alone. Given a set of trained *feature-MLPs*, a *merger-MLP* may be obtained from a training set in which each sentence is represented as the vector of *feature-MLP* outputs. This two level architecture was primarily chosen because it facilitates the analysis of the contribution of each features, by sampling the performance of the *feature-MLPs*.

We investigated a large set of candidate features, which could be divided into four categories: position of the sentence in the story, length of the sentence, similarity of the sentence to the overall document, and distribution of named entities (NEs) within the sentence. After some preliminary experiments, we settled on the set of eight features listed in table 1. The first three features can be classified as style features, and are concerned with length and position. The remaining features concern the content of the sentence. TF.IDF I and COSINE I are based on traditional information retrieval term weights comprising information about *tf* (*term frequency*) and *idf* (*inverse document frequency*) [12]. COSINE I is the cosine similarity measure of the *tf.idf* term vector to the document term vector. The final three features all concern the NE distribution in the sentence.

Feature	Description
POSITION I	Reciprocal position from the start.
POSITION II	Sentence position from the start.
LENGTH I	Length of sentence in words.
TF.IDF I	Mean of normalised *tf.idf* terms.
COSINE I	Cosine similarity measure of *tf.idf* terms.
NE I	Number of NEs.
NE II	Number of first occurrences of NEs.
NE III	Proportion of different NEs to number of NEs.

Table 1. Description of sentence-level features. The 'start' and 'end' are relative to the boundaries of the particular news story topic. NE = named entity. Counts of NEs are per sentence. The normalised *tf.idf* features, TF.IDF I are calculated as follows: TF.IDF I $= \frac{1}{\#words} \sum_w \frac{tfidf_w}{\sqrt{\sum_{w'} tfids_{w'}}}$.

For the text data NE annotations from the DUC evaluations have been used. The speech data was processed by an automatic NE recogniser [13].

2.3 Results

We assessed the contribution of an individual feature by basing a summariser on the relevant *feature-MLP* alone. Figure 2 shows the ROC curves[6] for four of the single feature summarisers and a summariser combining the whole feature set; each operating on both text and speech data. For both text and speech the summariser based on the full feature set had the best performance characteristics. For text, the positional feature POSITION I is clearly the most informative for summarisation; for speech there is no similarly dominant feature. This is linked to the stylistic differences between print and broadcast media.

These stylistic differences are also reflected in the contribution of the last style feature, the length feature (LENGTH I). For text, the sentence length is of less importance, but for speech it contains a lot of discriminative information about whether a sentence is summary-worthy. In the speech domain, the high information regions in the stories are often from the anchor in the studio, the main reporter or the occasional expert. It is often well-formed speech with longer sentences (either read or partly scripted speech). In contrast short sentences tend to be less information-rich.

The conclusions are similar when looking at the other main group of features, the content features. In text none of these features have been able to compete with the simple, yet very effective position features. In the speech domain, the content features contribute significantly. A very noticeable difference is for the

[6] An ROC curve depicts the relation between the false negative and true positive rates as the classifier output threshold varies.

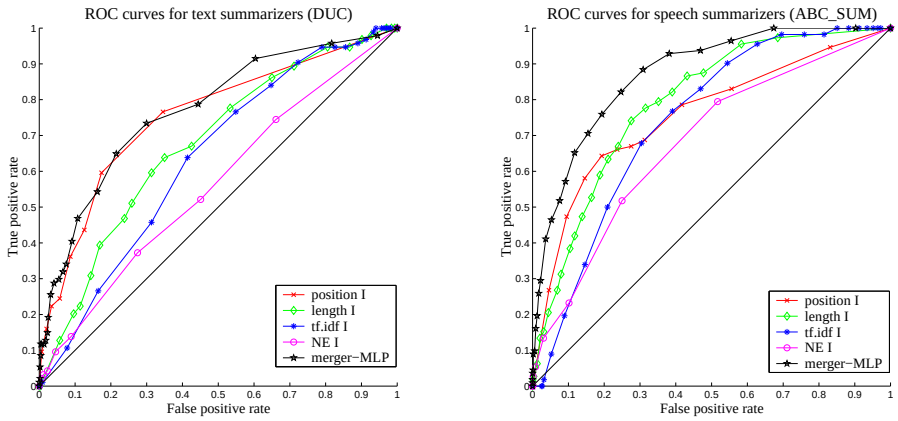

Fig. 2. Influence of the various features on the text and speech summarisers - ROC curves for the individual features and their combination to newspaper summarisation (DUC; left) and broadcast news summarisation (ABC_SUM; right).

named entity based features. Their performances in the text domain are relatively poor, but again the uneven information distribution in speech means that named entities become much stronger indicators of fact filled sentences. The *tf.idf* based features tell much the same story.

A final point to note is that for text the combination of the complete eight features added only minimal improvement to the performance of the best single feature summariser—based on the simple position feature. In the speech domain, the single feature summarisers are more complementary and their combination is significantly better than any of them alone.

Although the newspaper text and the broadcast news speech data are chosen so as to be as closely matched as possible, the gold-standard summaries differ: multi-line summaries for the text and one-sentence summaries for the speech. This discrepancy between data sets adds a level of complexity when drawing conclusions from these experiments. In terms of the contribution of the individual features it is likely that the apparent lack of contribution from some of the content features on the text data is partly down to the fact that when creating a multi-line summary any sentence candidate must not only be high in information relevant to the content of the story but also be a complementary match to sentences already selected in the summary.

The above experiments on broadcast news were carried out on manual, closed-caption transcriptions. Although these transcripts are not error-free (WER of 14.5%) they are still far better than transcripts from ASR systems. However, applications for automatic summarisation of spoken news stories would have to make do with transcripts output from automatic speech recognisers. Figure 3 shows the ROC curves for speech summarisers based on transcripts from six different ASR systems (produced for the TREC–8 SDR evaluation), along with the manual transcript. Each summariser was trained and tested on transcripts

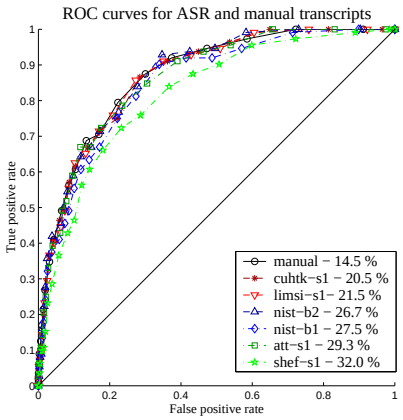

Fig. 3. . The influence of various WERs on the speech data summarisers - ROC curves for summarisers corresponding to different quality ASR transcripts plus the closed-caption transcript [14].

from the same source. The results indicate that there is relatively little difference due to WER, although the summariser based on the recogniser with the highest WER does show some degradation in performance.

3 Information extraction on Spontaneous and Read news stories

The observed relative indifference to WER is similar to that observed in spoken document retrieval using this data [15], and can be explained, at least in part, by the structure of a typical broadcast news story. The most information rich regions of a broadcast news story usually correspond to planned studio speech; spontaneous speech in variable acoustic environments is less information rich, from the point of view of summarisation—and harder to recognise. Zechner *et al.* report an increase in summarisation accuracy and a decrease in WER on broadcast news summaries by taking into account the confidence score output by the ASR system when producing the summary, and thereby weighting down regions of speech with potentially high WERs [16]. Kikuchi *et al.* propose a method that in an initial stage removes sentences with low recognition accuracy and/or low significance [17].

Clearly, factors such as the structure of the news story, the WER of the transcript and the types of feature do have an effect on the summary. For ABC_SUM, the structure of the news stories varies: some have a diffuse spread of information, others are more reminiscent of newspaper stories.

Here we report experiments that aim to quantify the effect of using traditional text features (content and style based) on automatically generated transcripts of SPONTANEOUS and READ news stories respectively. Experiments were performed

using three sets of transcripts from the ABC_SUM collection: the closed-caption transcripts (WER = 14.5 %), transcripts from the Cambridge University HTK[7] ASR system (WER = 20.5 %) and transcripts from the Sheffield/Cambridge Abbot[8] ASR system (WER = 32.0 %).

The news stories were manually classified into SPONTANEOUS and READ stories according to the following two categories:

- **Spontaneous news**: This category includes in it all the news stories which have both planned content and spontaneous utterances made by possibly multiple subjects apart from the news-reader. Typically this category includes street interviews, question/answer based conversations and large group discussions, but may also include interviews and individual discussions.
- **Read news**: This category incorporates all the news stories whose content is pre-planned and contains no spontaneous utterance. Usually these news stories tend to be short in length compared to the other categories. Typical examples for this category are financial reports and weather reports.

These categories represent a refinement of the classification applied in [18].

3.1 Results

Figure 4 shows four plots arising from doing summarisation on SPONTANEOUS and READ news stories based on high WER, low WER and closed-caption transcripts. Each plot shows ROC curves from four typical *feature-MLP* summarisers as well as from the *merger-MLP* combining all eight features.

Comparing plots for the SPONTANEOUS and READ stories (left-hand column to right-hand column) shows that the different types of feature perform differently depending on the style or category of the news story. On the SPONTANEOUS stories the position feature is much less important than for the READ stories. The sentence length and the *tf.idf* based features, on the other hand, are far more important in the SPONTANEOUS stories.

Only subtle differences in summarisation accuracy arise from an increasing WER. The curves for the closed-caption and low WER transcripts are very similar. For the SPONTANEOUS/high WER combination the area under the ROC curves is smaller, reflecting the increased number of errors in the transcripts. A larger difference is observed for the READ/high WER stories where the length and content based features have dropped in performance, in contrast to the position feature (which is not directly dependent on the speech recogniser).

These experiments further confirm the observed link between the feature contribution and the structure of a news story, and is in line with the conclusions drawn in section 2. In our previous work, the manual classifications into SPONTANEOUS and READ were not available and instead we performed an automatic classification based on the length of the news story [9]. The results are rather similar for both cases.

[7] The cuhtk-s1 system in the 1999 TREC–8 SDR evaluation.
[8] The shef-s1 system in the 1999 TREC–8 SDR evaluation.

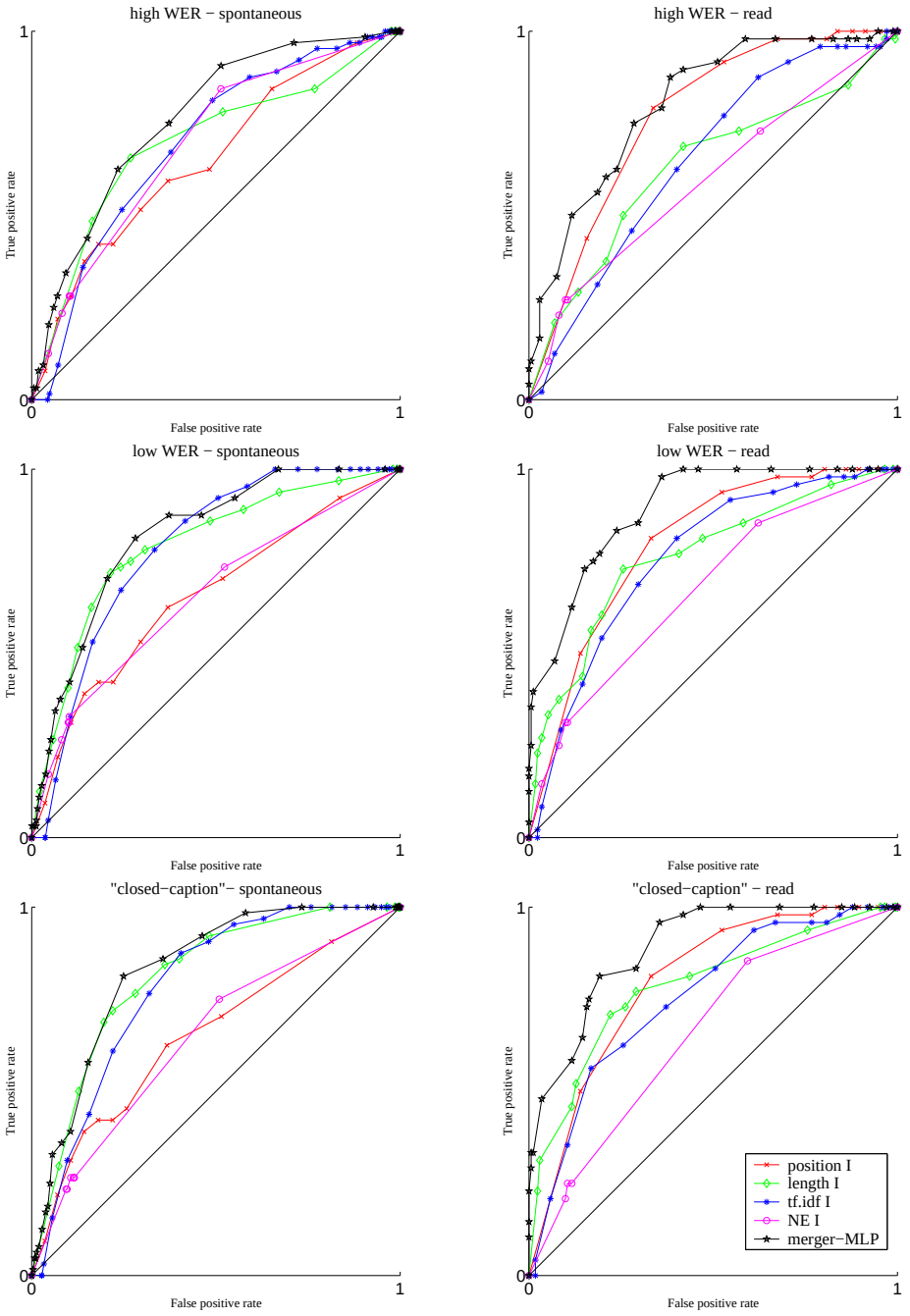

Fig. 4. The performance of summarisers based on all features and on four typical single feature summarisers on SPONTANEOUS and READ news stories and high WER, low WER and closed-caption transcripts.

4 Investigating style-specific summarisation approaches

The previous experiments have shown that the optimal choice of features when transferring text features to broadcast news is dependent on both the structure of the news story and the quality of the transcripts.

The experiments reported in this section explore the hypothesis that the observed difference in information spread in a spoken news story can be exploited, resulting in style-specific summarisation approaches.

We have used three different sentence extractive summarisers to automate this operation. The first incorporates a novelty factor to extract sentences for a summary, using an iterative technique that groups sentences which are similar to the document, but dissimilar to the partially constructed summary. The second selects the first line of the document (assumed to be a summary lead) and those sentences within the document that are similar to the first sentence. The third picks up the whole chunk of text around the sentence that is most similar to the document as a whole. For all the three summarisers we apply *tf.idf* weighting and re-arrange the selected sentences in the order of their appearance in the original document.

4.1 Data

These experiments are carried out using a portion of the hand transcripts from the Hub–4 acoustic model training data [19]. The transcripts are not case-sensitive and are devoid of any punctuation, such as sentence boundaries. For the work reported here, we manually split each segment of the transcriptions into individual news stories and marked the sentence boundaries.

For this data multi-line, gold-standard summaries were extracted by humans. The experiments were evaluated by a set of human judges, who scored the gold-standard summaries as well as three summaries obtained from the novelty, content and context based automatic summarisers.

4.2 Summarisation approaches

Summariser using novelty factor. This summariser is based on the maximum marginal relevance (MMR) algorithm [2] proposed by Carbonell and Goldstein, and builds an extractive summary sentence-by-sentence, combining relevance (similarity to the document) with a novelty factor (dissimilarity to the partially constructed summary). At the k^{th} iteration, it chooses

$$s_k \equiv \hat{s} = \underset{s_i \epsilon D/E}{\operatorname{argmax}} \left\{ \lambda Sim(D, s_i) - (1-\lambda) \max_{s_j \epsilon E} Sim(s_i, s_j) \right\} \tag{1}$$

where s_i is a sentence in the document, D is the document and E is the set of sentences already selected in the summary. D/E gives us the set difference, sentences not already selected. To form the summary the selected sentences, $\{s_k\}$ are re-arranged in the appearance order of the original news story. Sim is

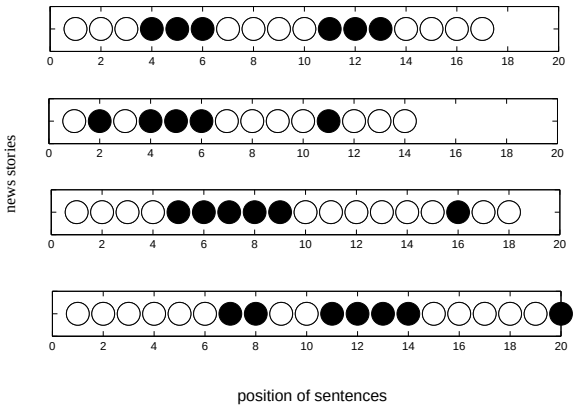

Fig. 5. Illustration showing the occurrence of sentences which are included in the summary for a given document. Each row represents a document (news story) and the sentence is represented by a circle. Each filled circle in the graph implies the sentence chosen by the human summariser to be included in the summary.

the cosine similarity measure. The constant λ is the weight of the novelty factor. $\lambda = 0.65$ was selected for experiments in this paper, based on some preliminary experiments on another database (BBC news transcripts).

Summariser using content. It is well-established that the first line of a textual news story is often a summary-worthy sentence, and this holds for some broadcast news stories. For example, our experiments have indicated that the first sentence is included in a human-generated extractive summary for about two-thirds of broadcast news stories. We can use this observation to design a summariser that extracts the first sentence, and treats it as a seed, extracting those other sentences that are most similar to it. The summary is a re-arrangement of $\{s_k\}$ that are selected by

$$s_k \equiv \hat{s} = \underset{s_i \epsilon D/E}{\operatorname{argmax}} \{Sim(s_1, s_i)\} \tag{2}$$

Sim is the cosine similarity measure.

Summariser using context. Another feature of extractive summarisation is that highly relevant sentences tend to occur in clusters. This is illustrated in figure 5 which shows which sentences were chosen to form part of the summary (extracted) by a human. The third summariser is based on this observation, with the sentence that is most similar to the whole document being chosen as a seed:

$$\hat{s} = \underset{s_i \epsilon D}{\operatorname{argmax}} \{Sim(D, s_i)\}, \tag{3}$$

categories of news stories	number of	sentences			words		
	documents	min	avg	max	min	avg	max
SPONTANEOUS	15	23	32	64	361	650	1562
READ	7	16	27	48	272	570	969

Table 2. Statistics of the 22 documents (news stories) used.

with the summary being formed by choosing those sentences adjacent to this seed sentence, $\hat{s}$. The summary is thus the seed sentence and its context.

4.3 Results

Each news story was manually classified into one of the two categories defined in section 3, and four summaries (three automatic, one human) were generated. Their quality was then evaluated by human judges.

We selected 22 news stories from the corpus, which were classified into the two categories. While the READ category had 7 news stories, the SPONTANEOUS news stories had 15 news stories and they varied in terms of size (Table 2). Each news story was summarised using each of the three automatic summarisers (novelty, content and context). The summarisers grouped the sentences forming a third of document or 100 words, whichever was larger.

As a benchmark, corresponding gold-standard summaries were generated by native English speakers. For the sake of uniformity of evaluation, the human summarisers were instructed to select the sentences from the document which they would ideally include in a summary, in the order of appearance.

The four summaries for each document were then rated by a set of four human judges (different from the people who summarised the documents) using a 1–10 scale, where 10 was the best. In order to obtain inter-judge agreement on the summariser, we have calculated κ [20], defined by

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{4}$$

where $P(A)$ is the proportion of the times that the l judges agree and $P(E)$ is the proportion of the times we would expect the l judges to agree by chance. Given that we are looking at l judges evaluating N document/summary pairs out of a score of a maximum of M for each category, we had to calculate the κ for each category. $P(A)$ and $P(E)$ are defined as

$$P(A) = \left[\frac{1}{Nl(l-1)} \sum_{i=1}^{N} \sum_{j=1}^{M} n_{ij}^2 \right] - \frac{1}{l-1} \tag{5}$$

where n_{ij} is the number of judges agreeing on a score of j for i^{th} summary.

$$P(E) = \sum_{j=1}^{M} p_j^2 \tag{6}$$

Summariser	Human	Novelty	Content	Context
κ	0.49	0.41	0.39	0.52

Table 3. Agreement among four judges for evaluation of various summarisers.

where p_j is proportion of the summaries assigned a score of j. If there is complete agreement then $\kappa = 1$ else if there is no agreement among the k raters $\kappa = 0$. The judges are said to be in moderate agreement when the κ is about 0.4 to 0.6. Table 3 shows the κ values for the four judges, indicating a moderate level of agreement.

The results of the human evaluations of the four summarisers in both the categories are shown in figure 6. Each scores for each summariser in the graph are averaged scores over the number of stories in that category.

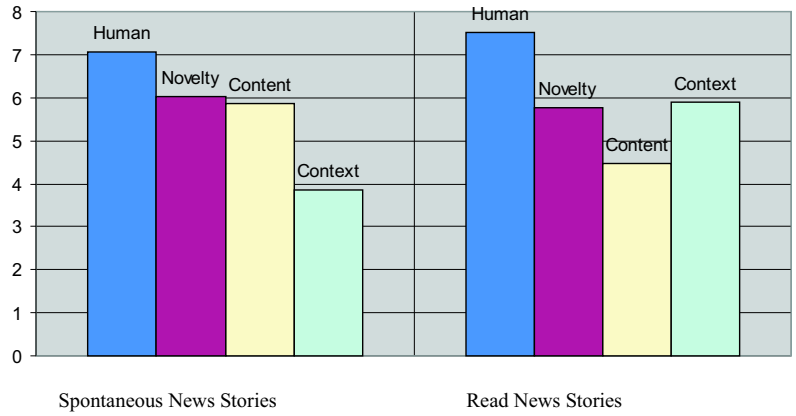

Fig. 6. The performance of the four summarisers on both the categories of news stories.

The human summaries were judged to be the best for 18 out 22 stories, with the largest deviations in the respective ratings occurring in the SPONTANEOUS news story category.

The automatic summarisers using novelty and content performed similar to each other for SPONTANEOUS news stories, and they both achieved better performance than the context-based summariser. For READ news stories, the context-based summariser performs best on some stories, the novelty-based summariser is best on others; on only one READ news story was the content-based summariser the best. The insignificance of the first line in some read news stories (see figure 5), especially in weather reports and financial reports, leads to a drop in performance of the content summariser on READ news stories.

The context-based summariser performs better than the other two summarisers, on the READ news stories category, which has a higher density of information than the SPONTANEOUS news stories. Its performance degrades on spontaneous news stories or stories with a high degree of data sparseness. The judges pointed out that this summariser fails for this category of broadcast news stories as it fails to highlight the real issues of the document.

The novelty- and content-based summarisers tended to lack coherence, with phenomena such as unexplained subject-object references and dangling anaphora[9]. This problem is avoided by the context-based summariser, which produces more coherent summaries, but at the cost of occasional repetition.

5 Conclusions

We have investigated the portability of extractive text summarisation techniques to broadcast news. We assessed the contribution of individual features (stylistic and content-based) by investigating ROC curves for summarisers based on newspaper data and broadcast news data respectively. It was found that for text the position feature is very dominating, and features containing content information are less important. For speech however, the stylistic features and the content features were all significant.

We have shown that classical text summarisation features are largely portable to the domain of broadcast news. However, the experiments reported here also made evident that the different characteristics of a broadcast news story, such as the different information distribution and the effect of different types of transcript error, warrant more sophisticated information extraction techniques, where the organisation of summary-worthy information in the news story is more explicitly taken into consideration.

Indeed we found that different summarisers may be appropriate to different styles of news story, particularly considering whether the presentation consists of planned or spontaneous speech. The novelty-based and content-based summarisers perform well on the classes with a spontaneous element. Context-based summarisation techniques are mainly limited to planned content.

We have demonstrated that different summarisation approaches clearly have their strengths and weaknesses, which should be expoited in relation to different categories of news stories.

References

1. S. Furui, "From read speech recognition to spontaneous speech understanding," in *the Sixth Natural Language Processing Pacific Rim Symposium, Hitotsubashi Memorial Hall, National Center of Sciences*, Tokyo, Japan, 2001.
2. J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries.," in *Proceedings of SIGIR*, 1998.

[9] For example, *"Another reason why. . ."* in the summary without the mention of first reason, *"and it ended tragically. . ."* without mentioning what "it" was and so on.

3. H. P. Edmundson, "New methods in automatic extracting," *Journal of the ACM*, vol. 16, no. 2, pp. 264–285, 1969.
4. S. Teufel, *Argumentative Zoning: Information Extraction from Scientific Articles*, Ph.D. thesis, University of Edinburgh, 1999.
5. S. Maskey and J. Hirschberg, "Automatic speech summarization of broadcast news using structural features," in *Proceedings of Eurospeech 2003*, Geneva, Switzerland, Sept. 2003.
6. C. Cieri, D. Graff, and M. Liberman, "The TDT-2 text and speech corpus," in *Proceedings of DARPA Broadcast News Workshop*, 1999.
7. J. Garofolo, G. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of RIAO-2000*, Apr. 2000.
8. J. Gauvain, "The LIMSI SDR system for TREC," in *Proceedings of TREC-9*, 2000.
9. H. Christensen, Y. Gotoh, B. Kolluru, and S. Renals, "Are extractive text summarisation techniques portable to broadcast news?," in *Proceedings of ASRU2003*, St. Thomas, US, Dec. 2003.
10. C. M. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, England, 1995.
11. J. Kupiec, J. O. Pedersen, and F. Chen, "A trainable document summarizer," in *Proceesings of SIGIR*, 1995.
12. C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 2001.
13. Y. Gotoh and S. Renals, "Information extraction from broadcast news," *Philosophical Transactions of the Royal Society of London, series A*, vol. 358, pp. 1295–1310, Apr. 2000.
14. S. E. Johnson, P. Jourlin, K. Spärck Jones, and P. C. Woodland, "Spoken Document Retrieval for TREC-8 at Cambridge University," in *Proceedings of TREC-8*, 2000.
15. S. Renals, D. Abberley, D. Kirby, and T. Robinson, "Indexing and retrieval of broadcast news," *Speech Communication*, vol. 32, pp. 5–20, 2000.
16. K. Zechner and A. Waibel, "Minimizing word error rate in textual summaries of spoken language," in *Proceedings of NAACL-ANLP-2000*, Seattle, WA, May 2000.
17. T. Kikuchi, S. Furui, and C. Hori, "Automatic speech summarization based on sentence extraction and compaction," in *Proceedings of ICASSP*, Hong Kong, 2003.
18. B. Kolluru, H. Christensen, Y. Gotoh, and S. Renals, "Exploring the style-technique interaction in extractive summarization of broadcast news," in *Proceedings of ASRU2003*, St. Thomas, US, Dec. 2003.
19. "Focus conditions for broadcast news evaluation, hub4," http://www.nist.gov/speech/tests/ bnr/ hub4_96/h4spec.htm, 1996.
20. S. Siegel and N. J. Castellan Jr, *NonParametric Statistics for the Behavioral Sciences*, McGraw-Hill International Editions, 1988.